ISSN: 0976-3104

SPECIAL ISSUE (SCMDSA)
Sree et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

ANALYSIS OF THE APPLICATION OF DATA MINING TECHNIQUES IN THE FIELD OF EDUCATION

Kaitha Sai Sree*, BMurali Manohar, Swarnalatha P

VIT University, Vellore, Tamil Nadu, INDIA

ABSTRACT

In olden days students exposed to teaching and learning was confined only to classroom. Students in the 21st century are connected well with lot of literature and learning from various open source platforms. Hence the students find it difficult to identify the right content for their learning. In this context data mining tools which are evolved in the past decade are of great use in filtering the large volumes of data and to pick the appropriate content for the student. In the previous years a considerable amount of time has been spent on student profiles and not on the factors responsible for the varied performances of the students. Data mining plays an important role in this field and can be used to mine relevant data for further optimization. This paper explains about how various techniques of educational data mining can be used to identify the student profiles and their behavior on the social media by taking many factors into consideration, to forecast the students' performance and also identify the best suited curriculum structure for them, to understand the pitfalls in teaching-learning environment etc. All these predictions can be used to help the 'at risk' student's category. The aim of the paper is to contribute to the literature on the application of data mining in the field of education by providing an overview of the various mining techniques that can be used, challenges faced in implementing the techniques and a comparative study of all the techniques.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26th-March-2016
Published on: 10th–June-2016

KEY WORDS

Knowledge Discovery in Databases (KDD); K-Means Clustering; Apriori Algorithm; Educational Data Mining (EDM); Feature Selection

***Corresponding author:** **Email:** kaitha.saisree2013@vit.ac.in; **Tel:** +91-8608556133

INTRODUCTION

Large volumes of data are added in every field of life on the internet by people across the globe through various platforms. It becomes a challenge for any person to identify the relevant information. The selection of useful information from the huge amount of data is known as Data Mining. Data Mining also refers to the application of algorithms for extracting patterns from data without the additional steps of Knowledge Discovery in Database (KDD) process. The results of this can be used to predict the most contributing attributes towards academia by making the good use of the enlarging amount of data in the field of education. Educational Data Mining (EDM) is concerned with developing new methods and techniques to understand the students ability of learning.[1]Due to the use of more computer based learning , the amount of data being generated has increased to a great extent , an interest to develop techniques for analyzing this data has also increased. These techniques are very useful in raising the educational standards and managements. It helps us to understand how people can become skilled in educational sector. Data Mining has many techniques like classification, Clustering, Association, Decision Tree, Neural Networks etc and the results obtained will help the educators to redesign their curriculum and to improve their learning methods, used to discover patterns which characterize learners across groups based on their choice of study and goals, used to attract the desired set of students, used to maintain student base and prevent dropouts by identifying students at risk and taking necessary actions etc. This paper deliberates about the various methods and algorithms that can be used to mine the data for educational purposes.

MATERIALS AND METHODS

Education is the process of learning, development of knowledge to enlighten oneself. The methods of education are teaching, discussing, training etc. Educational technology is the use of technological methods to improve the learning experience for the learners and the educators. These technologies have been changing and improving every day from many years. Initially a wooden board with lessons was used as the teaching medium, later after many decades the black board and chalk was introduced. This has been a means to teach the students for many hundreds of years. Distance learning was most preferred by

THE IIOAB JOURNAL
ISSN: 0976-3104

the people before the PCs were introduced into the education system. In the early 20th century TV (television) were used as means of teaching , later in the mid-20th century the PC(Personal Computer) was introduced and the same is used as a means of teaching. Later many gadgets were developed for the ease of the learners like the calculator etc. By the end of the 20th century WWW(world wide web) came into existence.

Distance-Learning and E-Learning

Distance education or the distance learning is a means of learning where the students are not physically present in the classroom. The teaching can be done using the internet or it can partly have classes on site which is known as courses conducted through partly distance learning. Students were provided with the material by the institution which offers the distance learning programs. For clarification the student can reach through the student counselor in person at a specified place. With the advent of technology i.e. Internet, E-learning etc. the learning experience is brought into the door steps of aspirants to gain knowledge and enhance skills by various open source platforms like coursera, function space etc. These platforms provide a experiment , to interact with course provide by raising questions through discussion forum. Data mining plays a critical role for customising the courses according to the students learning pace and convenience. This results in customisation of learning modules rather than standardised modules.

Web Mining

The experience of distance learning is improved by applying the mining techniques to the Wold Wide Web(WWW) which is known as web mining.With the amount of data increasing on the internet , mining of this data can help us discover interesting patterns in distance learning and can help in customizing the courses for the students.

In the past many decades, the educational practices have hardly been changed. With the implementation of the distance learning there are some major changes that will take place like the relationship between the student and the teacher, the teacher will be a coach now and will lead the group of students i.e. the coach teaches the students and keeps track of their performance and gives feedback. The students will be more independent and tend to learn them and discuss with the peers and helps them develop individually. They make the actual use of the technology like the computer, internet etc.

These distance education platforms like udacity, coursera etc., gather large volumes of information which is generated automatically by the web servers and are collected and stored in the servers. Some other sources of the information can be taken based on the references the students browse. Such data can be helpful for the institution to improve the structure of the learning process by understanding the learning patterns of the students. [2]

The mining techniques like clustering can be applied to group students based on the skills they possess or other criteria. Other techniques can be used to predict the performance of the students. For example, if a student has attempted many practice tests these results can be used to predict the performance of the students in the final tests and estimate their performance to check for the areas which needs improvement and work on those areas and generate rules. By applying these rules, if we can find simple association rules it can helps us in improving the structure.

The data collected from the students can be improved further by asking them about their prior experiences and collect more extensive data and apply the mining techniques on that data to discover interesting patterns about the student behavior and improve the courses offered accordingly and customise them for the students. [3]

This web based learning methods collect a lot of information on user patterns etc. Data mining techniques can be applied on this data to predict the final grade of the students, and faculties can use this information to concentrate more on students who need help and advise them accordingly. This will be of more use in classes that have many people.

The data generated in the field of education has drastically increased after the E-Learning has been introduced. The data generated through this is very huge and plays a very important role if converted into knowledge in improving the learning experience for the learners and the educators. The sources of data include, data generated from the educational settings like, from the universities, intelligent tutoring systems etc. At a higher level, the data in this field needs to be explored in a proper way in order to discover new perceptions on how the data generated can be used.

E-Learning (Electronic Learning) refers to learning conducted via electronic media, generally the internet; it also refers to intentional use of networked information, communication technology information and communication technology in teaching and learning. This can be one form of distance learning, and it involves the use of internet to download materials, interact with the instructor etc. [4]

Common Attributes of E-Learning System

The common attributes for a E-Learning System can be identified in many levels and sections.[5]
Some common attributes of E-Learning System are mentioned in the following table [**Table-1**].

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 1. Common Attributes of E-Learning System

S.No	Attributes	Description
1	Visited	If the unit, document or web page has been visited
2	Total_time	Time taken by the student to complete the unit
3	Score	Average final score for the unit
4	Knowledge_level	Student's initial and final level in the unit
5	Difficulty_level	Difficulty level of the unit
6	Attempts	No. of attempts before passing the unit
7	Chat_messages	No. of messages sent/read in the chat room
8	Forum_messages	Number of messages sent/read in the forum

Procedure for selecting E-Learning Resources

Step by step procedure to select E-Learning Resources

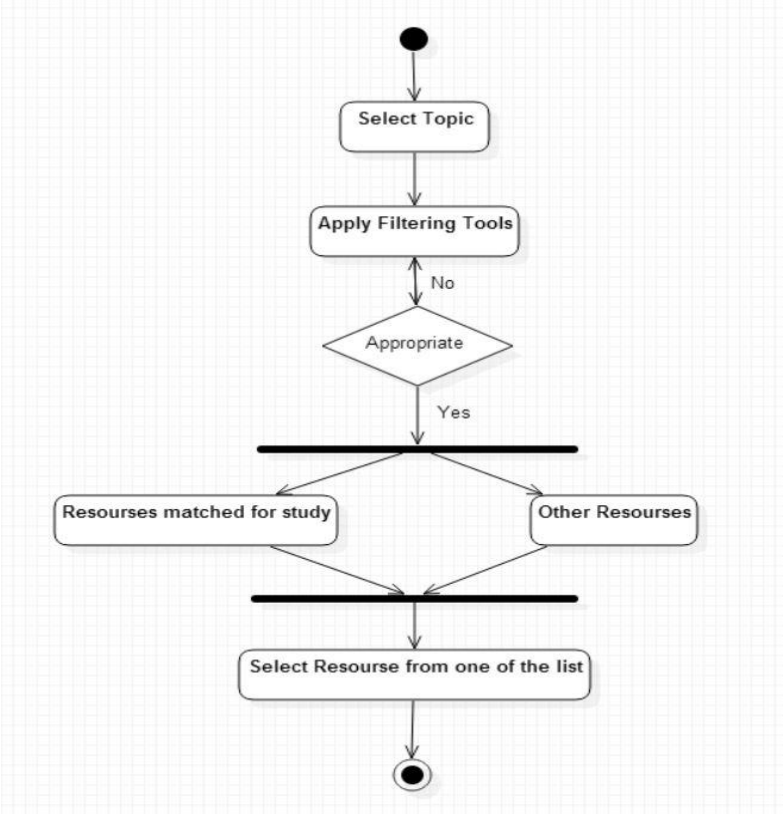
1.**Filtering:** The students are first asked to select a topic of choice, then some filtration rules are applied on the topic. Later depending on the results the topic is declared to be appropriate or not appropriate for the student.

2.**Prediction**: According to the topic chosen other related results are also predicted and shown.

3.**Decision**: Two rank lists are made one based on the topics that are suggested for the student to study and the other based on the knowledge level of the students and difficulty level of the topic.

4. **Adaptation**: After the lists are made, the students select a topic from one of the lists. [6]

The steps involved in selecting E-Learning resources is shown in the following flow chart in **Figure-1**



\Fig: 1. Step by step procedure to select E-Learning Resources

……

THE IIOAB JOURNAL
ISSN: 0976-3104

E-learning can be of great advantage to students from rural areas to take up many courses which aren't available due to lack of resources like such as lack of qualified faculty, infrastructure in their academic institutions. Many working executives won't find time to go to college to take up additional skill development, certification courses so they make use these online portals to complete their courses. Many courses are offered by global institutions across the world where anyone interested can take the course.

With the scope of E-Learning being increased, many e-learning platforms have been developed causing the amount of data generated to increase. This generated data plays a crucial role in optimizing the e-learning methodologies. The data produced can become the knowledge and help in improving the e-learning process in all the possible ways.

Feature Extraction Models

Feature extraction models are used to create new feature from the existing original features whereas feature selection models are used to select the best subset of features from the given feature set.

Feature selection models are generally used in the field of data mining to explain in detail about the tools and techniques that help to reduce large amount of data into manageable size for further processing and analysis. These models are developed with an aim to discover the combinations of attributes that have high classification accuracy. Complex data have a large number of variables. When large number of variables are involved the memory then the computation power required will be high and may cause classification algorithms to over fit. Hence this model is used to reduce the unnecessary data into reduced set of features.[7]

Among the popular feature extraction models RELIEF and FOCUS algorithms were found to give better combinations of feature subsets than other wrapper approaches. RELIEF algorithm assigns a relevant weight to each attribute of feature vector by computing difference between selected test instance nearest hit and nearest miss training instances. [8]

$$w_i=w_i\text{-}diff(x_i,near\text{-}hit_i)^2+diff(x_i,near\text{-}miss_i)^2 \quad (2.2.1)$$

The wrapper model finds out the subset using n-fold cross validation method in which the data set is portioned into equal sized partitions, where the unique partition is the data set and the rest n-1 partitions is the training data set.

With the help of the induction methods John et al. has extracted feature subsets using machine learning. They have researched about improving the accuracy of the prediction, it majorly used two versions namely the forward and backward version. The forward version started with all the features where as in the backward version the features weren't known at all. [9]

Martin Sewell did a survey on feature extraction which explained that subsets containing minimum dimensions contribute most to the accuracy. [10]

A model has been proposed by Micheal Fire to predict the success of a student in a course by taking the previous data related to the course and applying network analysis methods. Parameters like personal information, friends grades etc.

Azwa Abdul presented a survey to predict student performance focusing on three elements i.e. parameters, method and tools. The extraction of patterns has been done using the Naive Bayes Classifier. [11]

Feature selection is a commonly used preprocessing step in machine learning. It is categorized into filter and wrapper methods. In the filter method, from a set of all the features a best subset of features is selected, then the algorithm is learned and that results in performance. In the wrapper method from the set of all the features selecting a subset involves both generating a subset and learning algorithm and this result in performance. In wrapper methods the computation time is high when the number of variable is more.

The student data sets generally have his/her personal data, details of parent's education, occupation etc. and academic attributes (grade in high school, higher secondary etc.) tot behavioral attributes (gender, cast etc.).They also assess based on the evaluation parameters like total number of answers which were correct and the ones that were correct in the first attempt etc. The parameters considered in the case of a drop-out student were difficulties in grasping during the class, improper selection of academic preferences, financial factors, feminine problems etc. These prediction results are used in increasing the capability of the students during the period of their studies.
A method has been proposed by the author to predict student performance using historic patterns and help them by contributing to their efforts. This method has the following phases.

Phases in Predicting the Students Performance

Phase 1: Current Scenario of Attributes

Attributes influencing were gathered from various sources. These Attributes were of two types
1. External Attributes: These are changeable with academic effort **[Table-2]** 2. Internal/Inherent Attributes: These are not changeable with academic effort **[Table-3]**.

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 2. External Attributes

S.No	Variable	Explanation	Values
1	Attendance	Attendance of the student during the semester	Percentage
2	Assignment and Project marks	Marks scored out of 20	Marks on a scale of 20
3	Internal Score	Marks scored in 2 internals out of 30	Marks on a scale of 30
4	No of subjects	No of internal papers which the student has appeared for out of 7 registered courses	Number
5	Computer	Knowledge of using a computer	Low/Medium/High

Table: 3. Internal Attributes

S.No	Variable	Explanation	Values
1	Gender	Sex of the student	Male/Female/Other
2	Cast	Cast f the student	Gen/ST/SC/OBC
3	Medium	The medium of instruction	Hindi/English/Other Language
4	Food Habit	Food Behavior	Vegetarian/Non Vegetarian
5	Settlement area	Residence of the student	Village/Town/City
6	Residence	Accommodation in hostel or home	Yes/No
7	Class10 grade	Grades of 10^{th} class	92-100 - A1 83-91 - A2 75-82 - B1 67-74 - B2 59-66 - C1 51-58 - C2 43-50 - D1 35-42 - D2 <35 - E
8	Class 12 grade	Grades of 12^{th} class	91-100 - A1 81-90 - A2 71-80 - B1 61-70 - B2 51-60 - C1 41-50 - C2 33-40 - D 21-32 - E1 00-20 -E2
9	Fathers Qualification	Education of father	12^{th} /UG/PG
10	Mothers Qualification	Education of mother	12^{th} /UG/ PG
11	Family status	Type of family	Joint/Small
12	Fathers Occupation	Job of the father	Govt/Private
13	Mothers Occupation	Job of the mother	Govt/Private
14	Family Income	Yearly income of the family	<50000 50000-100000 100000-200000 200000-300000 300000-400000 400000-500000 >500000
15	Percentage of dropouts	No of dropouts in previous year	%

Phase 2: Preprocessing Of Data Sets

In this phase of the feature extraction model deal mostly with preprocessing of the data sets. Some of the inherent attributes like gender, location, medium of education also seem to have an effect on the results. So these are taken into consideration during prediction of "at risk" students.

Phase 3: Working Model

This phase involves the classification and the fitness evaluation tests with the criterion that the "at risk" category students will obtain less than 40% of the total score. By applying Naive Bayesian Classification the prediction is done.

THE IIOAB JOURNAL
ISSN: 0976-3104

Phase 4: Experimental Observations

In this phase we need to check how the internal and external factors affect the prediction. From the experiment it has been observed that external attribute i.e. students attendance was found to contribute the most for academic assessment which is then followed by internal assessment scores and assignment credit etc.

Results:

The objective of the experiment was to find out how the external attributes affect the academic performance of the students and the results of this prediction can help the management of the institution to take actions accordingly and improve their grades in upcoming examinations [8].

Clustering and Prediction

Prediction models are used to predict the event that is to occur in the future. It guesses the probability of an outcome based on the given input. Clustering is used to group the students according to similarities and this helps to improve the educational process and helps us in getting more specific models for the student's response. Artificial Neural Networks (ANN) is one of the best way to predict the performance of the student with a very simple and easy to use interface.

Student Performance Evaluation System (SPES)

Anju and Robin have done a research work on Decision tree classification algorithms like ID3 (Iterative Dichotomiser), C4.5 and CART were used to generate a decision tree from a data set, which is used to predict the performance of the student.ID3 accepts only categorical data whereas C4.5 which is the improved version of the ID3 accepts both categorical and continuous attribute and is highly efficient and accurate [12]..Ajai and Saurabh have proposed a method to extract knowledge using ID3, C4.5.Accurate result is not given by ID3 when there is noise and pruning is also not supported by ID3 whereas C4.5 is successful in identifying the students who are most likely to fail by using gain ratio [13].Kalpesh et al. discussed about the performance of students using ID3 and C4.5 i.e. the algorithms to generate the decision tree from data sets. They proposed and developed a system which can show the achievements of the students based on the previous performance[14].Brijesh Kumar et al has discussed about the decision tree methods for the classification which helps in identifying the dropout students and also the students who need special care and counseling[15].Narayana Swamy and Hanumanthappa have discussed about decision tree classification technique and predicted the academic success of the student enrollment which was very accurate[16].

The author suggested the various techniques which have been developed to see the data in different perspectives. The curriculum for the course is one of the main aspect which is required to improve the teaching of a particular course and make it more successful. The different perspectives suggested by the author can be used to develop a proper curriculum and there by improve the course. The author concludes saying that of the both algorithms ID3 and C4.5 which are used to identify the various categories of students , C4.5 algorithm is of higher superiority.

Apriori Algorithm and K-Means Clustering Algorithm

Apriori algorithm

Apriori algorithm is used to find out the frequent item sets in a transactional database. The principle of the algorithm says that the subset of a frequent item set must also be frequent. The most important use of this algorithm is in Market Basket Analysis.

Apriori algorithm majorly has 2 steps they are,

STEP 1: All the frequent item sets are found in this mode using the recursive mode.
STEP 2: The item sets which satisfy the confidence condition are acquired here.

Apriori algorithm is generally used on databases which contain transactions. Bottom up approach; breadth first search and Hash tree structure are used to count the candidate item sets appropriately. This is used when there is large item set property. It is easy to implement and can be paralleled easily. The main disadvantage of this algorithm is that it has to do many database scans and assumes that the database is memory resident. It is used in situations where we have the transactions of which students are taking particular subjects and the probability that they will take another subject as well in common can be found out using this algorithm.

K-Means Clustering Algorithm

Clustering is the process of partitioning the group of data point into small number of clusters. Lloyd's Algorithm also known as K-Means algorithm aims in partitioning 'N' observations or data objects into 'K' clusters in which each observation will belong to the cluster with nearest mean.
The center of the clusters is initialized first then the attribute that is closest to cluster of each data point. The position of the cluster is set to the mean of the data points which belong to the cluster, these steps are repeated until convergence.

This is used for classification of lot of information based on its own data. It is based on the comparison of numerical values that result the data based on the distance, hence it is considered as an unsupervised data mining algorithm. This will classify data into clusters and indicate which patterns belong to the class. This is the best method to be used, if the variables are huge and is

THE IIOAB JOURNAL
ISSN: 0976-3104

comparatively faster than other hierarchical methods. The main disadvantage of this algorithm is the difficulty in K value prediction. This is used in situations where we have to classify the students or related things into clusters accordingly.

These techniques are used not only to analyse the students but also be used by the faculties to see the students previous records and modify the course content accordingly for easy understanding of the students and this can be used by the school authorities to prepare a curriculum that is suitable for all the students [17].

Crisp DM Methodology (Cross Industry Standard Process for Data Mining)

Crisp DM Methodology is a well-made methodology that provides a structured approach to plan a data mining project. In this process many automated analysis techniques are used to extract the knowledge. This model is independent of the industry in which it is applied. This describes the different approaches the experts commonly use to deal with problems in data mining. The Methodology is a wider concept compared to the KDD (Knowledge Discovery in Database)
This methodology is mainly divided into 6 stages. The order in which the stages have to be performed is not particular.

1. Business Understanding

This phase focuses on understanding the objectives and the requirements of the project from the business point of view, and then converting the problem into data mining problem definition. In the step the objectives of the KDD should be set depending on the field of business for which the analysis is being conducted. E.g. the financial sector uses this to detect fraud, the retails industry for ensuring customer satisfaction etc. This can also be used in the education system in predicting the success of the student, optimization of the courses, attracting new students etc. To be up to date in the market one should understand the latest changes being made in the field and should know the changes in the customer profiles. Social Media should be used in the marketing strategy of the companies. From the results of the analysis on the data obtained from the social media, the desired group of customers should be targeted and the offers should be changed according to the customer preferences.

2. Data Understanding

This phase begins with collection of initial data in a suitable format followed by activities to get a clear understanding about the data so that the problems can be identified and to detect the hidden information.
The average grade of the student is affected by their behavior on the social media.
-The activities the students generally do when they are on the social media sites.
-Most of the students who use social media for educational purposes are successful.

- The survey done by the author broadly divided the questions into 3 categories.
- The first set was related to the general information from which the demographic information can be obtained.
- The second part is related to how familiar the respondents are to the social media and how frequent they visit the social media and the purpose of networking and sharing.
- Third they examine the use of social media for the selection of college and education needs.

Additional questions related to their activities on blogs, forums etc. and the reasons for not being active on the social media. The data collected is entered into a sheet for the easy understanding and manipulation. According to the analysis done on the data collected, it was observed that the respondents would prefer websites that were more flexible in communication and that offer a more entertaining way to communicate with people. The common reasons for using social media were found to be to connect with friends, watch and listen to videos, communicate with fellow students regarding materials etc. , to follow school activities , share videos , songs etc. and other conclusions included expressing views through social media.

3. Data Preparation

In this phase all the activities related to constructing the final data set from the initial data is covered. The tasks in this phase are performed multiple times without any particular order to be followed. The tasks include data selection, data cleaning, data construction, data integration and formatting data.
Many attributes are removed for some of the following reasons,

- Part of the data collected is used for statistical processing about the future students general information.
- Most of the data for a particular attribute has the same value, since this does not bring any knowledge it can be removed from the further analysis.
- For the missing data substitute the default or global value, or substitute with the mean etc. to complete the incomplete data. The data subsets are selected based on the data mining goal.

4. Modeling

In this phase different modeling techniques are selected and used for building the data mining model. The method selected depends on the problem that has to be solved. There are many solutions available to solve a particular problem. The techniques used depends on the staff available, quality of the data, it also depends on the time and other factors. For the purpose of clustering K-Means clustering can be used.
Grouping students based on activities they perform has resulted in four major groups,
Group 1: This group has the students who regularly check the comments in the forums and comment on posts, videos but are not the active creators of the content on internet.
Group 2: These groups of people are active creators of content on the group.

THE IIOAB JOURNAL
ISSN: 0976-3104

Group 3: These groups of students majorly use social media occasionally for the purpose of entertainment. They hardly leave any comments on the posts, videos etc. These students listen to songs, watch videos once in week approximately.
Group 4: Students in this often visit the blogs, forums but do not have the freedom to express their opinion on social media.

5. Evaluation

By this stage the model has been built and should be reviewed thoroughly before deploying the model. The main objective is to find if there are any issues that have to address.

6. Deployment

The project does not end with simple creation of the model. Depending on the requirements of the user, the model has to be deployed as needed by the user, it can be in the form of a simple report or can be as complicated as data mining process [18].

The results on using this methodology are, the essential step in this is to identify the various profiles of students as various users produce variety of groups and profiles. By understanding the behavior of the students on social media and internet there can be many advantages to this field as the data generated can be used for many purposes especially in publicizing about the institution. The results show that there are many distinct patterns among the students. There are students who use social media only for the purpose of entertainment and do not influence others whereas a portion of the students do get influenced by word of mouth which is also considered as publicity.

Data mining in higher education

Higher education is a way of learning provided by the universities through teaching, research work and other practical experiments. The quality of education provided by the universities depends on the faculties or the teaching staff that they recruit. The HR team of the university plays a very important role in selecting the members. The data mining techniques can be applied in this field to recruit the best staff. The mining techniques like clustering , classification , prediction play a very important roles in improving the performance of the organisation.In large organisations , performance evaluation is done on a yearly basis where the performance of each individual is done based on the qualities , abilities they possess. Based on the results of this test , the areas which need improvement can be figured , and the workforce is divided according to the knowledge they possess and assigns appropriate people to a particular job. This improves the overall performance of the education sector by achieving the organisational goals.

The performance evaluation of the employee is done taking all the factors that contribute to the performance into consideration. According to the general approach used the score of the employee is a simple number on given scale. But this score does not depict the various inter dependencies among the attributes used which could be more helpful in the performance determination. Few factors like responsibility, contribution in achieving the organisation goal etc. that contribute in measuring the performance are not directly measurable. Thus the performance evaluation is a tedious task if many attributes are considered to the evaluation. The author mainly focuses on the data mining techniques available, the model that supports both classification and prediction and its performance analysis.

According to the authors S.Anupama Kumar and M.N Vijayalakshmi , various mining techniques like classification and prediction can be applied on the student data to predict the performance of the students. This paper concludes that different methods or techniques have an advantage in different areas. [19]. M.Sukanya et al.the performance of the student can be improved in many ways by applying the mining techniques clustering, classification and prediction. The factors like psychological, social and personal play an important role in the performance of the student. By mining the information gathered can be useful to manage the next batch of students in a more efficient way and helps in improving the performance of the students [20].Data mining is a tool which helps on allocating the resources and the staff appropriately and helps in managing the resources. In higher education, the employees play a very important role and their management is very important for an effective performance. The performance of the employees is evaluated based on various contributing factors. The supervised and unsupervised learning techniques are used to build the models for performance prediction[21].Data Mining plays a very important role in Human Resources(HR) department as they can analyse the skills of the applicant and choose employees according to the need of the organisation[22].

With the changing methods of teaching and learning in higher education there are many issues that the field is facing. The employee should bear the role of being a mentor , educator , researcher etc. By considering only the teaching in classroom or the feedback given by the students which is biased , the performance of an employee cannot be predicted instead the performance should be predicted taking all the possible factors into consideration. About more than 50 factors like Attitude, Teaching skills, Communication skills etc. are considered in evaluating an employee. By evaluating all these attributes the performance of the employee is understood and are rated on a scale.[23]

THE IIOAB JOURNAL
ISSN: 0976-3104

Comparison of the Techniques/Algorithms/Methodologies

Table: 4. Comparison of various technologies

Technique/Algorithm/Methodology Used	When is this technique used?	Principle	Advantages	Disadvantages
Apriori Algorithm	This is used when we have to find the frequent item set from the given transactions	The subset of a frequent item set must also be frequent.	▪ Easy to implement. ▪ Easily paralleled. ▪ Uses large item set property.	▪ Assumes the transaction database to be memory resident. ▪ Requires many database scans.
K-Means Clustering Algorithm	This is used when we have to classify a lot of information based on its own data.	Partitions 'n' observations into 'k' clusters in which each observation belongs to the nearest mean.	▪ Computes faster than hierarchical clustering if k value is small. ▪ Produce tighter clusters than hierarchical clustering	▪ Difficult to predict the K-Value. ▪ Different initial partitions can result in different final clusters.
Feature Extraction Model	This is used when we have to reduce the amount of resources required to describe the large set of data.	When the input data is too large and is suspected to be unnecessary then it is transformed into reduced set of features.	▪ The memory and computation power required can be reduced. ▪ Has a very important application in image processing.	▪ Sometimes can be computationally expensive. ▪ They may fail to remove the redundant features.
Crisp DM Methodology	This model describes the commonly used approaches that experts use to tackle problems.	This methodology breaks the process of mining into six major phases namely Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment.	▪ Makes large data mining projects faster, cheaper, more reliable and more manageable. ▪ Even small scale data mining can be benefited. ▪ Provides unique framework for guidelines and experience documentation.	▪ There are many risks at every stage that have to be taken care off.
ID3 Algorithm	This is used when we have to construct a decision tree.	Entropy and Information Gain is used to select the most useful attribute for classification. Entropy is used to calculate the homogeneity of the sample. Information gain is based on the decrease in entropy after the data set is split on an attribute.	▪ Understandable prediction rules are created form the training data. ▪ Builds the fastest tree. ▪ Builds a short tree. ▪ Whole data set is searched to create a tree. ▪ Finding leaf nodes enables test data to be pruned, reducing number of tests.	▪ Data may be over-fitted or over classified if used on a small sample. ▪ Only one attribute at a time is tested to make a decision. ▪ Classifying continuous data may be expensive as many trees have to generate.
C4.5 Algorithm	This is used when we have to generate a decision tree.	The concept of information entropy is used in building the decision tree, i.e. It is based on the expected value of the information.	▪ Handle noisy data, missing data better. ▪ More memory efficient. ▪ Deals with continuous and discrete attributes. ▪ Easy to implement.	▪ Does not work well on small training set. ▪ Small variation in data can lead to different decision trees.

All the methods, algorithms, methodologies mentioned above are compared and presented in **[Table-4]**.This gives a clear understanding of the different methodologies available in the literature till date.

Visual Data Mining Techniques

The presentation of the data in a way that can be understood by the users is called as visualization. In the traditional method of evaluation response is obtained by checking out the previous works and collecting related materials. [24]This response plays a crucial role in the evaluation and shouldn't be ill-judged. The main aim of this visualization is to improve the feedback between the

THE IIOAB JOURNAL
ISSN: 0976-3104

user and graphs and to make it more interactive so that the system of evaluation can be more adjustable, dependable and user friendly.

A model has been proposed by the author, which follows a multi-tier architecture including the application layer, middle layer and the data layer. The main functions of this model are as follows.

Step1: The data is obtained by interactive means, and then it is sorted interactively and expressed in view.
Step2: The algorithms like decision trees, clustering, and association rules can be used according to the requirement. New algorithms can be added, if they support the format. Generally C++ is used to compile these algorithms as it is more stable.
Step3: The large amount of data is expressed so that the users can understand the data and can find out the hidden information. The visualization technologies used are
1.Column Graphs - This graph tells about the situation of an element at different points of time. The horizontal axis has the time and the vertical axis has the changing values. E.g. the performance of the students in different years can be depicted by this graph.
2.Scatter Graph - These graphs show how much one variable is affected by the other variable. The value of one variable shows the position on the horizontal axis, and the other variable gives the vertical axis position. Taking these two coordinates the point is marked on the graph. E.g. The relation between the capacity of the lung of a person and how long the person can hold his breadth.
3.Parallel coordinates - In this method of representation the n dimension space can be shown on a two dimension plane with n-parallel coordinate at equal distances.

By representing the data using these methods, the role of the users can be utilized to a maximum extent and unwise decisions can be reduced making good use of the knowledge possessed by the users [25].

CONCLUSION AND SCOPE FOR FUTURE STUDY

This paper has examined the concepts of data mining in the field of education and simultaneously gave the overview of the application of various mining techniques on how the data can be used in this field. Though a sizable amount of research is being carried out in this field , there is wide area which still has to be uncovered. Like the experimental data collected in many of the researches is just a small amount and the results obtained are only for that. To get a clear view these techniques have to be applied on huge amount of data collected intensively by interviews, questionnaires etc. so that the results can be more apt.

Researchers can focus on the application of data mining techniques in the field of education, so that it can help the educators to develop better curriculum's for the students. Apart from the internal and externals attributes mentioned many other parameters can be taken into consideration and used for the prediction of more detailed information about the student. This prediction can be used to advise the student on the choice of his/her major based on the parameters. They can also focus on the weighing the data mining techniques so that it can be useful to both the educators and learners. In the future the researches can collect extensive data by making questionnaire and then apply the techniques mentioned above to get the valuable results. The current research will require more in depth study to discover more valuable rules to get a clear picture of the applications of data mining in the field of education for its optimization .Furthermore this can be used to make groups among the students in the easier way, identify hidden patterns, find undesirable behavior of students etc. The results of all these can help the faculty to improve relationship with the students and help them reach out to students who are in need of their help. Data mining can be a part of the technologically advanced techniques. The authors would like to propose a hybrid algorithm in the future which will combine the existing clustering algorithms to reduce the gaps identified in the paper.

FINANCIAL DISCLOSURE
This research is self assisted financially.

ACKNOWLEDGEMENT
The authors thank VIT University, Tamil Nadu , India for providing all the facilities required.

CONFLICT OF INTERESTS
Authors declare no conflict of interest.

THE IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] RamandeepKaurBhinder el al. [2015] A Review on Using Cryptography Techniques for Securing User Data in Cloud Computing Environment. *International Journal of Computer Science & Communication (*IJCSC),.6:83–86.

[2] NiteenSurv et al. Framework for Client Side AES Encryption Techniques in Cloud Computing. *International Advance Computing Conference (IACC)*, 525– 528.

[3] Periyanatchi S, Chitra.K. [2015] Analysis on Data Security in Cloud Computing-A Survey. *International Conference on Computing and Intelligence Systems* 04:1281 – 1284.

[4] LovepreetKaur et al. [2015] A Survey on the Encryption Algorithms in the Cloud Security Applications. International journal of Science Technology & Management (IJSTM), pp.1– 9.

[5] Neha A Puri et al. [2014] Deployment of Application on Cloud and Enhanced Data Security in Cloud Computing using ECC Algorithm. 1667– 1671.

[6] Thiyagarajan B, Kamalakannan R. [2014] Data Integrity and Security in Cloud Environment Using AES Algorithm. Information communication and Embedded Systems. 1– 5.

[7] CharanjeetKaur et al. [2015] Data Security Algorithms In Cloud Computing: A Review. *International Journal For Technological Research In Engineering* 2:372– 375.

[8] Sana Belguith et al. [2015] Enhancing Data Security in Cloud Computing Using a Lightweight Cryptographic Algorithm. *The Eleventh International Conference On Autonomic and Systems*. 98– 103.

[9] Tembhurne S et al. [2015] An Improvement In Cloud Data Security That Uses Data Mining. *International Journal of Advanced Research in Computer Engineering & Technology* 4: 2044– 2049.

[10] Nikhitha K, Navin K S. [2015] A Survey On Various Encryption Techniques For Enhancing Data Security In Cloud. *International Journal of Advanced Research Trends in Engineering and Technology* 194– 197.

[11] Rashmi S, et al. [2015] Architecture for Data Security In Multi-cloud Using AES-256 Encryption Algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication* 157-161.

[12] Masthanamma V, et al. [2015] An Efficient Data Security in Cloud Computing Using the RSA Encryption Process Algorithm. *International Journal of Innovation Research in Science, Engineering and Technology* 4: 1441– 1445.

[13] SaiSindhuTheja R et al. [2015] Data Security in Cloud for Medical Sciences using AES 512-bit Algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication* 1746– 1749.

[14] Nasrin K, ZurinaMohd. [2014] A Framework Based on RSA and AES Encryption Algorithms for Cloud Computing Services. *IEEE Conference on Systems, process and Control* pp. 58-62.

[15] Sugumaran M et al. [2014] An Architecture for Data Security in Cloud Computing. 2014 World Congress on Computing and Communication Technologies. pp. 252– 255.

[16] Arockiam L, Monikandan. [2014] Efficient Cloud Storage Confidentiality to Ensure Data Security. *International Conference on Computing Communication and Information*. 5:1– 5.

[17] Anuj Kumar et al. [2014] Cloud Data Security using Authentication and Encryption Technique. *International Journal of Innovative Research In Technology* 1: 388– 391.

[18] Vanya divan et al [2014] Cloud security solution: comparison among various cryptographic Algorithms. *International journal of advanced research in computer science and software Engineering* 4:1146– 1148.

[19] pradeep Kumar et al [2014] An authentication approach for data sharing in cloud environment for dynamic group. International conferences on issues and challenges in intelligent computing techniques(ICICT)9:262–267.

[20] Meenakshi et al. [2014] Data security analysis in cloud environment. *International journal of innovations & advancement in computer science* 2:14– 19.

[21] Aized Amin Soofi et al [2014] Encryption Techniques for cloud data confidentiality. International Journal of Grid Distribution computing. 7:11–20. .

[22] Vishwanth, S.Mahalle et al [2014] Enhanced the data security in cloud by implementing hybrid(RSA&AES)encryption algorithms. *International conference on automation and communication*. pp.146–149.

[23] Prashantrewagad et al. [2013] Use of digital signature with Diffie Hellman key Exchange and AES encryption algorithm to enhanced data security in cloud computing. *International conference on communication systems and network technologies*. 3:437-439.

[24] Ching-Nung yang, jia-bin lai. [2013] Protecting data privacy and security for cloud computing based on secret sharing. *International symposium on biometrics and security technologies* 7:259–266.

[25] ParsiKaplana ,sudha. [2012] Data security in cloud computing using RSA algorithm. *International journal of research in computer and communication technology*, vol.1.

[26] Rohit ,sunil [2012] A proposed secure framework for safe data transmission in private cloud. *International journal of recent technology and engineering*, vol.1

[27] Aleksandar et al. [2012] Data confidentiality using fragmentation in cloud computing. *International journal of networks and distributed system*, 1:85–90.

[28] Mohand M et al [2012] Enhanced data security model for cloud computing. International conference on Informatics and systems. vol.36, pp.cc-12.

[29] PachipalaYellamma et al. [2013] Data Security In Cloud Using RSA. 4'th International Conference on Computing Communication and Networking Technologies, pp. 1-6.

[30] Sonia sindhu. [2015] A survey of security algorithms in cloud computing. International journal of Advanced Research in Computer Engineering & Technology,.4(5):2368–2371.

[31] Ramesh K, Ramesh S. [2014] Implementing One time password based security mechanism for securing personal health records in cloud. International Conference on control, Instrumentation, Communication and computation technologies. pp. 968–972.

[32] Subbhiah S, Selva S [2015] Distributed data security for data prevention in cloud computing using One time password for user authentication. Journal of Environmental Science, *Computer Science and Engineering & Technology* 4:752–758.

[33] Priyanka Nema [2014] An Innovative Approach for dynamic Authentication in Public cloud: Using RSA, Improved OTP and MD5. *International Journal of Innovative Research in Computer and Communication Engineering*. 1(11):6697–6702.

[34] RandeepKaur, SupriyaKinger. [2014] Analysis of Security Algorithms in Cloud Computing. *International Journal of Application or Innovation in Engineering & Management* 3: 171–176

THE IIOAB JOURNAL
ISSN: 0976-3104

[35] SunithaSharma et al. [2013] Enhancing Data Security In Cloud Storage. *International Journal of Advanced Research in Computer and Communication Engineering* , 2: 2132–2134

[36] Vijendra et al. [2014] Data Storage Security in Cloud Environment with Encryption and Cryptographic Techniques. *International Journal of Application or Innovation in Engineering & Management*, 3: 209–213

[37] Jay Singh et al. [2012] Improving Stored Data Security In Cloud Using RC5 Algorithm. Nirma University International Conference on Engineering. pp. 1–5.

[38] DeepikaVerma, Karan Mahajan. [2014] To Enhance Data Security in Cloud Computing Using Combination of Encryption Algorithms, 2: 41–44.

[39] Honey Patel, JasminJha. [2012] Securing Data in Cloud Using Homomorphic Encryption. *International Journal of Science and Research*. 4, :1892–1895.

[40] Jayanthi M et al. [2014] Analysis on Secure Data Sharing using ELGamal's Cryptosystem in Cloud. *International Journal of Computer Science and Electronics Engineering*, 4:50–55.

[41] Raghul et al. [2015] Data Security in Federated Cloud Environment using Homomorphic Encryption Technique. *International Journal of Emerging Technology and Advanced Engineering*, 5:137–141.

[42] Vishal Paranjape, VimmiPandey [2013] An Approach towards Security in Private Cloud Using OTP. *International Journal of Emerging Technology and Advanced Engineering* 3:.683–687.

[43] Abhishektripathy, TarunGoyal. [2014] Cloud Data Security Using Encrypted Digital Signature & 3D Framework. *International Academic of Science, Engineering and Technology* 3:114–121.

[44] ShikhaChoksi. [2014] Comparative Study on Authentication Schemes for Cloud Computing. *International Journal of Engineering Development and Research*, 2: 2785–2788.

[45] HanumanthaRao et al. [2013] Data Security in Cloud using Hybrid Encryption and Decryption. *International Journal of Advanced Research in Computer Science and Software Engineering*. 3: 494–497.

[46] Roshani et al. [2015] Data Security in Cloud through Confidentiality and Authentication. *International Journal for Scientific Research & Development* 3: 1735–1738.

[47] Dimpi Rani, Rajiv. [2014] Enhanced Data Security of Private Cloud Using Encryption Scheme with RBAC. *International Journal of Advanced Research in Computer and Communication Engineering* 3: 7330–7337.

[48] Pradeep et al. [2012] Enhancing Data Security in Cloud Computing Using 3D Framework & Digital Signature with Encryption. International Journal of Engineering Research & Technology. 1:1–8.

[49] Ranu S, Hasna. [2015] Biometric Based Approach for Data Sharing in Public Cloud. *International Journal of Advanced Research in Computer and Communication Engineering*. 4:95-97.

[50] SuchitaKolhe et al. [2015] Five-Level Authentication Security in Cloud Computing. *International Journal for Research in Emerging Science and Technology*, 2:116–118.

[51] Sasi E, Saranyapriyadharshini.[2015] Secured Biometric Authentication In Cloud Sharing System. *International Journal of Computer Science and Mobile Computing*, 4: 572–577.

[52] Sudhansu & Biswaranjan [2014] Enhanced data security in cloud computing using RSA encryption and MD5 Algorithm. International *Journal of Computer Science Trends and Technology* 2(2):60–64.

ABOUT AUTHORS

Ms. Kaitha Sai Sree *is a 4th year , M.Tech Software Engineering student , in the School Of Information Technology , VIT University , Vellore.She interned at two major stat ups Axinovate Technologies and Think Tankers Innovative Solutions at Hyderabad where she worked on Back-end technologies and Mobile App development.Her current research interest include Data Mining , Data Analytics and Big Data.*

Dr. B. Murali Manohar, Senior Professor - VIT Business School*, V I T University.He was a Visiting Professor for six months at CMIS, University of Cologne, Germany sponsored by DAAD. He completed the UNDP assignment for a period of 2 years- employed by Ministry of Education, Govt. of Ethiopia.at Debub University.he is actively involved in the activities of NBA accreditation, ISO –9002 from DNV, Netherlands and Deemed University status from the ministry of HRD, Govt. of India at VIT. He pursued his Ph. D in* ***E-Commerce.****He is having 2 years of Industrial experience and more than 20 years of teaching experience both at UG and PG level Management Programs.Received Fellowship from Rotary Int, U S A to visit East Yorkshire, U K.He published more than 65 papers in International Journals/International Conference Proceedings/National Conferences. He has reviewed 3 Management books and published the same.Two research scholars have completed their Ph. D under his guidance and five more Ph.D scholars are pursuing their research.Currently a reviewer/member for many of the reputed International/National Journals.*

THE IIOAB JOURNAL
ISSN: 0976-3104

Swarnalatha Purushotham *is an Associate Professor, in the School of Computing Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D degree in Image Processing and Intelligent Systems. She has published more than 57 papers in International Journals/International Conference Proceedings/National Conferences. She is having 15+ years of teaching experiences. She is a senior member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), ACEEE.She is an Editorial board member/reviewer of reputed International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.*